**Feasibility 1: Limited Data on r/altright**

We pulled the dataset of unique users who posted on r/altright in October and November of 2016 (Election season). There were approximately 3,000 users initially. Given that the subreddit had around 8,000 subscribers before being banned, this isn’t entirely surprising. However, we then filtered for users who made (1) more than 5 posts in r/altright during that time period, as well as (2) filtered for people who had posts older than a year prior to their first r/altright post Unfortunately, this left us with only 384 users. While their post histories are long (with more than 100 users with accounts more than 4 years old) and rich, this still poses an issue of whether enough users exist on the subreddit. To account for this, we have deiced to look at numerous subreddits of related (e.g. r/the\_donald) and opposing (e.g. SandersForPresident) viewpoints.

**Feasibility 2: Data Download Issues**

At the current moment, we can pull just less than 16,000 rows at a time from Google BigQuery. While this is fine for pulling unique author lists, this poses a challenge when pulling a large subreddit’s posts, as well as the histories of thousands of users. With the set of 384, 120 different csv files were produced through manual querying. While an API exists to greatly simplify this process, it seems that it only allows pulling off of projects that *you own*. This can be simplified by copying over queried tables from u/fhoffa’s original dataset. However, this also requires storing and querying your own data – which requires money. Another solution is to torrent u/stuck\_in\_the\_matrix’s data (approx. 250GB) – which is much cheaper, but much more of a pain to futz with.

**Feasibility 3: Theoretical Plausibility**

We had a chance to look through the 384 users and what subreddits they participated in during various time intervals prior to their first r/altright post. This could potentially clue us in onto how similar users are to the r/altright without needing to do any language yet. Intiially, similar users are to the r/altright without needing to do any language yet. Initially, looking at the top ranked subs (most posted in our sample over time) aren’t entirely surprise – as they consist primarily of default subs (r/askreddit, r/WTF, etc.) However, if we remove any subreddit from the ranking who that ever been a default sub, we find that similar subreddits to r/altright were visited throughout the entirety of the time frame. This provides some evidence for H1a, though we need the language model to confirm these results more robustly.

<Link>

**Feasibility 4: Building Language Models**

Using Jack’s script (only needing to edit it for Python3), we were able to successfully compare two corpuses and the distance between them easily. The only issue we have is how to translate the resulting list of tuples into a singular measure of distance.

<Link>